Array-Ready Oligo Set™ for the Mouse Genome Oligo Set Version 2

We are pleased to announce Version 2 of the Mouse Genome Oligo Set, containing 16,463 70mer probes representing 16,463 genes. Using refined probe design methods, we have formulated a new set of probes for most known mouse genes. This set represents many new genes not included in our Mouse Genome Oligo Set Version 1.1, and incorporates more sophisticated design methods along with more gene sequence information. We also provide a map from the Mouse Genome Oligo Set Version 1.1 that corresponds to this new set. An amino linker is attached to the 5' end of each oligo.

Gene sequence source and selection

All probes are designed from the UniGene Database Build Mm 102 (February 2002) and the Mouse Reference Sequence (RefSeq) Database, both developed and maintained at the National Center of Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov).

Advantages of using gene sequences from UniGene and RefSeq

UniGene is an open source database freely available to everyone. UniGene automatically clusters GenBank sequences into a nonredundant set of gene-oriented clusters. It has become one of the most widely used *de facto* standards in the public domain for cataloguing mouse genes. Each UniGene cluster contains cloned genes and expressed sequence tag (EST) sequences that represent a unique gene. UniGene sequences are filtered for contaminant sequences, genomic repetitive regions, and low-complexity sequences using NCBI's Dust program. All oligos are designed from the representative sequence of each of the 16,463 clusters. The representative sequence chosen is the sequence with the longest region of high-quality sequence in each cluster.

The NCBI Reference Sequence (RefSeq) project is an effort to standardize gene sequence references by providing an NCBI-staff curated gene sequence and avoiding gene sequence redundancy. Each RefSeq is linked through a LocusID number to NCBI's LocusLink Database. RefSeqs genes are quickly becoming the *de facto* standard and are supported by a large research community.

The advantages of designing a probe directly from a gene sequence from the RefSeq collection are given below. Each of these advantages is provided by the NCBI LocusLink Interface.

Features provided for each RefSeq	Example
Alternate gene symbols	VPF, VEGF-A
Full-length coding regions known	Present
Gene Ontology™ (GO) terms	Biological process: Cell cycle control
Chromosome/Cytogenetic map information	17 24.2 cM
Official gene symbol and name by Mouse Genome Database (MGD)	Vascular endothelial growth factor (VEGF)

In Version 2 of the Mouse Genome Oligo Set, 6951 oligos are designed directly from a gene sequence from RefSeq.

Number of oligos designed using information from UniGene and RefSeq	6951
Number of oligos designed from UniGene	9512
Total number oligos in Mouse Genome Oligo Set Version 2	16,463

Probe design and selection rules

Once a gene has been selected to be included in the set, a probe is selected with an optimal set of parameters. Large numbers of 70mer candidate probes for each gene are selected using the following criteria for the Mouse Genome Oligo Set.

1) All oligos are within 78°C ±5°C using the following formula:

   $T_m = 81.5 + 16.6 \times \log[Na+] + 41 \times (\#G + \#C)/\text{length} - 500/\text{length}$

   where $[Na+] = 0.1$ M and $\text{length} = \#A + \#C + \#G + \#T$

2) Each oligo is within 1000 bases from the 3' end of the available gene sequence.

3) An oligo cannot have a contiguous single nucleotide repeat or poly (N) tract longer than 7 bases.

4) An oligo cannot have a potential hairpin structure with a stem length longer than 9 bases.

5) A normalized score is assigned to each oligo based on the number of repeats. Oligos with more repeats having a normalized score greater than a certain threshold are filtered out.

6) Each oligo has less than or equal to 70% identity to all other genes. For all oligos in the Mouse Genome Oligo Set Version 2, using BLAST, each oligo is aligned

against all 87,550 representative sequences in Mouse UniGene Build Mm102. Using the alignment with the candidate oligo versus the highest scoring non-self gene, a BLAST percent identity score is computed. The highest scoring non-self gene is defined as the sequence that yields the most matched bases in an alignment. This BLAST percent identity is also referred to as cross-hybridization homology or similarity of the oligo. This calculated percent identity score is dependent on the size of the sequence database used to BLAST against, oligo sequence, and use of either gapped or no-gap alignment method.

7) Each oligo of any length cannot have greater than 20 contiguous bases common to any other gene.

Once oligo candidates have been selected satisfying all the selection rules mentioned above, each oligo is ranked based on BLAST percent identity as computed in Step 6. One final oligo for each gene is selected with the minimum percent identity or cross-hybridization similarity.

For a small number of genes that did not yield oligos satisfying all the above criteria certain rules were relaxed. For those genes, the oligo is selected anywhere in its sequence or is designed to be less than 70 bases long.

SUMMARY

Oligo selection criteria	Value	Number of oligos in genome set satisfying these criteria
Length Melting temperature Location from 3' end Poly(N)tract length Stem length in potential hairpin Cross-hybridization to all other genes Contiguous base match to any other gene	70mer 78°C ±5°C <=1000 <=7 <=9 <=70% <=20	**16,245**
Length Melting temperature Location from 3' end Poly(N)tract length Stem length in potential hairpin Cross-hybridization to all other genes Contiguous base match to any other gene	70mer 78°C ±5°C Any <=7 <=9 <=70% <=20	**218**
Total		**16,463**

The following illustrations show the distribution of all 16,463 oligos for melting temperature, GC content, location from 3' end of gene sequence, length of maximum stem length, and BLAST percent identity or cross-hybridization similarity.

Figure 1. Melting Temperature

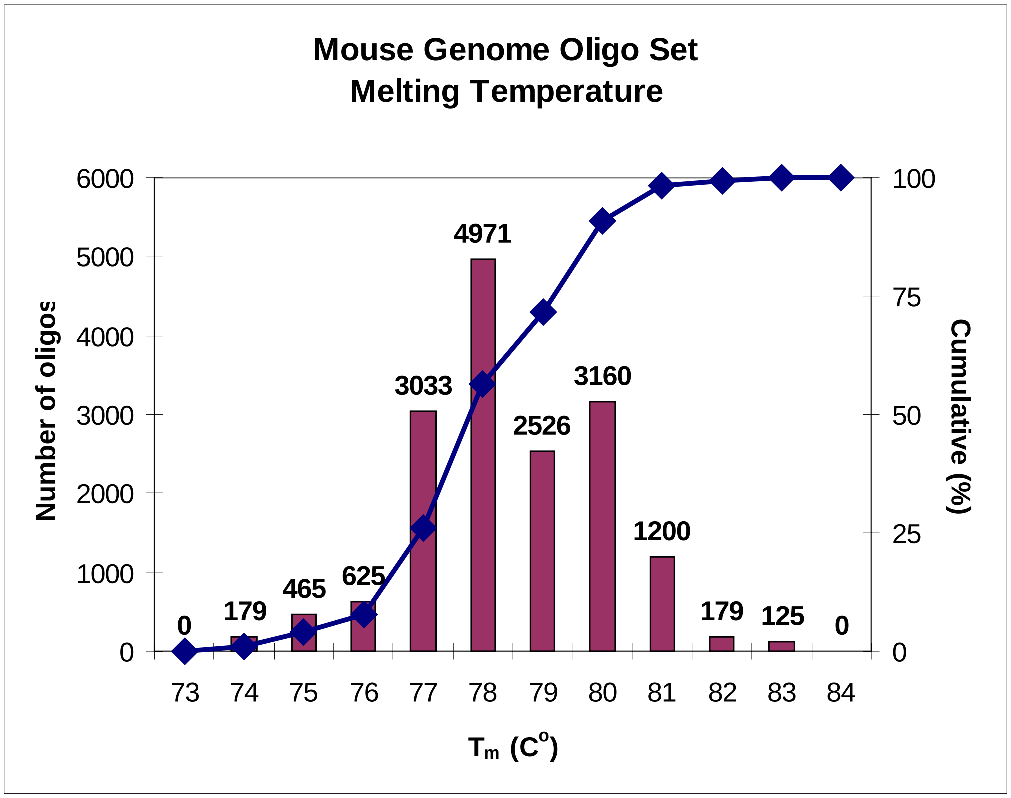


Figure 2. GC Content

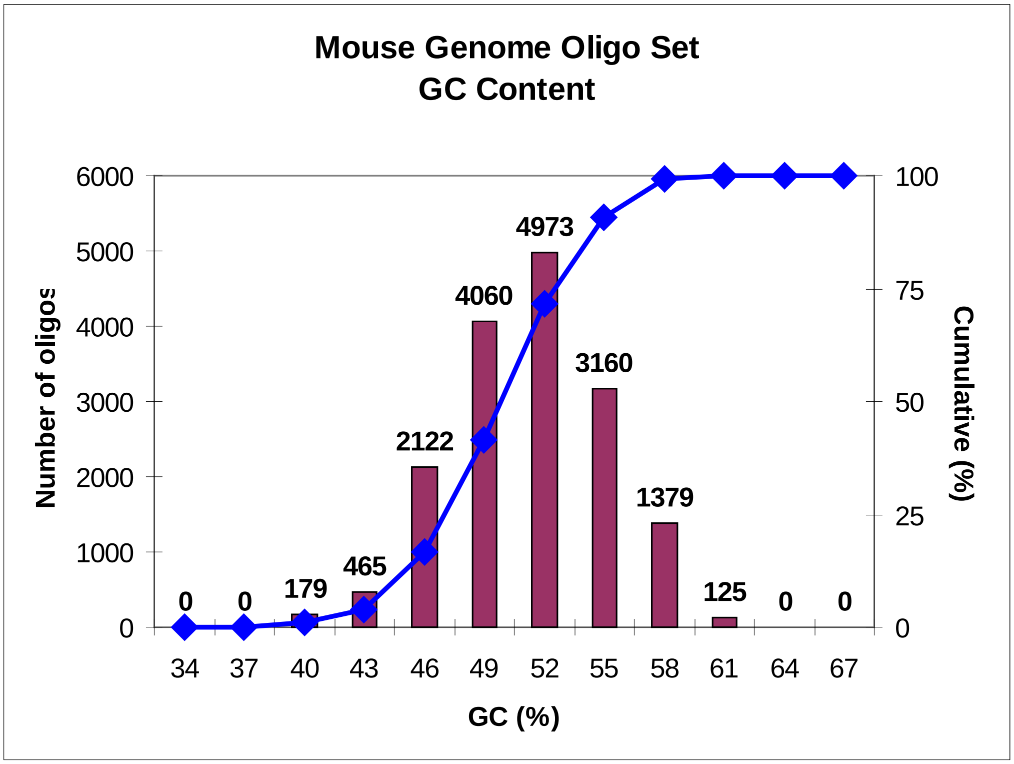


Figure 3. Location from 3' End

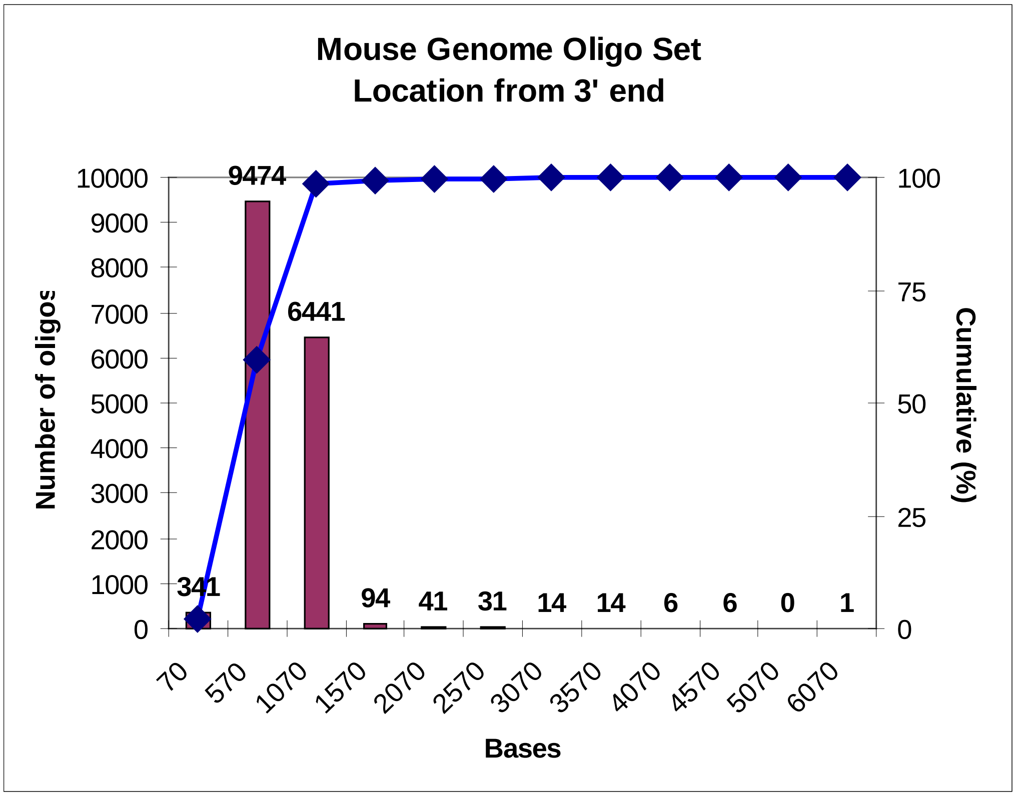


Figure 4. Length of the Longest Hairpin Stem

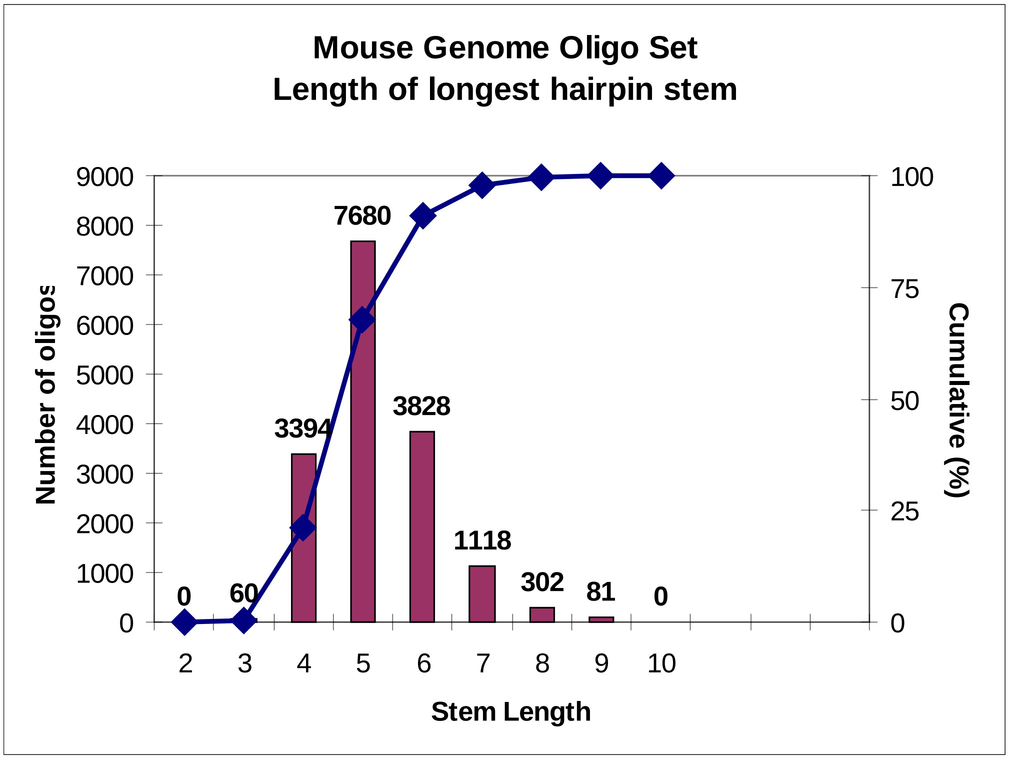
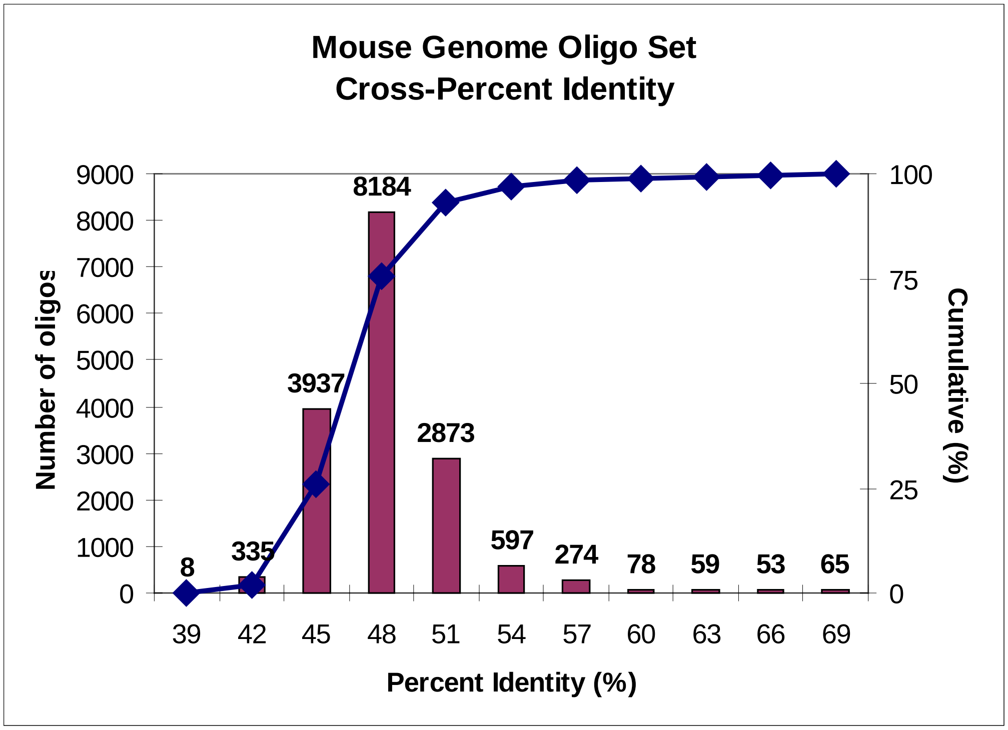


Figure 5. Cross-Hybridization Identity



Quality check of probe design specifications

Once the final oligo has been selected to represent a gene, each oligo undergoes design specifications quality control where we use an independent method to confirm that all oligos have met the specified design specifications. The table below summarizes data from our quality check for probe design specifications for all 16463 oligos in the set.

Probe Design Specification	Expected Value	Verified Range
Melting temperature (C°)	78°C ±5°C	73.1–82.61
GC content (%)	35–70	37.7–60.1
Hairpin stem length (base pairs)	< = 9	3–9
Cross-hybridization similarity (%)	< = 70	37–69